Data Mining as a Tool for Environmental Scientists

Jessica Spate [a], Karina Gibert[b], Miquel Sànchez-Marrè[c], Eibe Frank[d], Joaquim Comas[e], Ioannis Athanasiadis[f], Rebecca Letcher[g]

[a]Mathematical Sciences Institute, Australian National University, Canberra, Australia

[b]Department of Statistics and Operation Research, Technical University of Catalonia, Barcelona, Catalonia

[c]Knowledge Engineering and Machine Learning Group, Technical University of Catalonia, Barcelona, Catalonia

[d]Department of Computer Science, University of Waikato, Waikato, New Zealand

[e]Laboratory of Chemical and Environmental Engineering (LEQUIA), University of Girona, Girona, Catalonia

[f]Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, Lugano, Switzerland

[g]Integrated Catchment Assessment and Management Centre, Australian National University, Canberra, Australia

Abstract:

Over recent years a huge library of data mining algorithms has been developed to tackle a variety of problems in fields such as medical imaging and network traffic analysis. Many of these techniques are far more flexible than more classical modelling approaches and could be usefully applied to data-rich environmental problems. Certain techniques such as Artificial Neural Networks, Clustering, Case-Based Reasoning and more recently Bayesian Decision Networks have found application in environmental modelling while other methods, for example classification and association rule extraction, have not yet been taken up on any wide scale. We propose that these and other data mining techniques could be usefully applied to difficult problems in the field. This paper introduces several data mining concepts and briefly discusses their application to environmental modelling, where data may be sparse, incomplete, or heterogenous.

Keywords: Data mining; environmental data.

1 INTRODUCTION

In 1989, the first *Workshop on Knowledge Discovery from Data (KDD)* was held. Seven years later, in the proceedings of the first *International Conference on KDD*, Fayyad gave one of the most well known definitions of what is termed *Knowledge Discovery from Data*:

> *"The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in*

data" Fayyad et al. [1996b]

KDD quickly gained strength as an interdisciplinary research field where a combination of advanced techniques from Statistics, Artificial Intelligence, Information Systems, Visualization and new algorithms are used to face the knowledge acquisition from huge data bases. The term *Knowledge Discovery from Data* appeared in 1989 referring to high level applications which include particular methods of *Data Mining*:

> *"[...] overall process of finding and interpreting patterns from data, typically interactive and iterative, involving repeated application of specific data mining methods or algorithms and the interpretation of the patterns generated by these algorithms"* Fayyad et al. [1996a]

Thus, *KDD* is the high level process combining *Data Mining* methods with different tools for extracting *knowledge* from data. The basic steps established by Fayyad are briefly described bellow. In Fayyad et al. [1996a] details on the different techniques involved in this process are provided:

- Developing and understanding the domain, capturing relevant prior knowledge and the goals of the end-user
- Creating the target data set by selecting a proper set of variables or data samples (including generation of proper queries to a central data warehouse if needed)
- Data cleaning and preprocessing. Quality of result is dependent on the quality of input data, and therefore the preprocessing step is crucial. See Section 3.1 for discussion of this point
- Data reduction and projection: Depending on the problem, it may be convenient to simplify the set of variables in question. The aim here is to keep a relevant set of variables describing the system adequately and efficiently
- Choosing the data mining task, with reference to the goal of the KDD process. From clustering to time series forecasting, many different techniques exist for different purposes, or with different requirements. See Section 3 for a survey of the most common data mining techniques. Depending on the choice of methods, various parameters may or may not need to be set, with or without optimization
- Selecting the data mining algorithm/s: once decided the task and goals are codified, a concrete method (or set of methods) needs to be chosen for searching patterns in the data. Depending on the choice of techniques, parameter optimization may or may not be required
- Data mining: Searching for patterns in data. Results from this stage will be significantly improved if previous steps were performed carefully
- Interpreting mined patterns, possibly followed by further iteration of previous steps
- Consolidating discovered knowledge: documenting and reporting results, or using them inside the target system.

The steps outlined above can be illustrated as Figure 1 (from Fayyad et al. [1996a]). Fayyad's proposal, outlined above, marked the beginning of a new paradigm in KDD research:

> *"Most previous work on KDD has focussed on [...] data mining step. However, the other steps are of considerable importance for the successful application of KDD in practice"* Fayyad et al. [1996a]

Fayyad's proposal included prior and posterior analysis tasks as well as the application of data mining algorithms. These may in fact require great effort when dealing with real applications. Data cleaning, transformation, selection of data mining techniques and optimization of parameters (if required) are often time consuming and difficult, mainly because the approaches taken should be tailored to each specific application, and human interaction is required. Once those tasks have been accomplished, the application of data mining algorithms becomes trivial and can be automated, requiring a only a small proportion of the time devoted to the whole KDD process. Interpretation of results is also often time consuming and requires much human guidance.

However, it is common in some scientific contexts to use the term *Data Mining* to refer to the whole KDD process Siebes [1996] instead of the application to a cleaned dataset only. In those contexts the

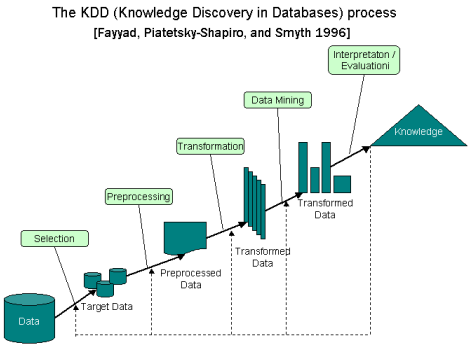

Figure 1: Outline of the Knowledge Discovery Process

process may be thought of, in brief, as a sequence of four main steps: data cleaning and variable selection, algorithm and parameters selection, application of said algorithm, and interpretation of results. Some research attention has recently been given to the data mining 'process model' (Shearer [2000]), where an addition phase of *deployment* is also discussed.

It is clear that either referring to the knowledge discovery process as *KDD* or simply as *Data Mining*, tasks like data cleaning, variable selection, interpretation of results, and even the reporting phase are of much importance as the data analysis stage itself. This is particularly true when dealing with environmental data, which is often highly uncertain, and may contain outliers and missing values that necessitate special treatment from any modelling scheme.

Selected algorithms are discussed in Section 3, along with preprocessing methods, which will be explored in Sections 3.1 to 3.3. Later in the paper, concerns such as performance evaluation, model optimization and validation, and dealing with disparate data sources, are dealt with. Section 4 contains a brief review of previous environmental data mining work and Section 6 a discussion of some existing environmental problems that may particularly benefit from data mining. Available software is discussed in Section 7, with particular reference to the Weka (Whitten and Frank [1991]) and GESCONDA (Sànchez-Marrè et al. [2004]) packages. The role of iEMSs in encouraging data mining is raised in Section 8.

2 Some Notes on Environmental and Natural Systems and Modelling

Environmental systems typically contain many interrelated components and processes, which may be biological, physical, geological, climatic, chemical, or social. Whenever we attempt to analyse environmental systems and associated problems, we are immediately confronted with complexity stemming from various sources:

Multidisciplinarity: a variety of technical, economical, ecological and social factors are at play. Integration of different social and scientific disciplines is necessary for proper treatment, as well as use of analysis techniques from different scientific fields

Ill-structured and nonlinear domain: environmental systems are poor or ill structured domains. That is, they are difficult to clearly formulate with a mathematical theory or deterministic model due to high complexity. Many interactions between animal, vegetal, human and climatic system components are highly nonlinear

High dimensionality and multiscalarity: most environmental processes take place in two or three spatial dimensions, and may also involve a time component. Within this frame, multiple factors are acting at many different spatial and temporal scales (see Section 3.2)

Heterogeneity of data: for environmental real world problem, data comes from numerous sources, with different formats, resolutions and qualities. Qualitative and subjective information is often integral. This topic is discussed in more detail in Section 6.1

Intrinsic non-stationarity: Environmental processes are in general not static, but evolve over time. The assumption of stationarity cannot be justified in many physical, chemical, and biological processes Guariso and Werthner [1989].

Controllability: Controllability of environmental systems is poor, due the unavailability of actuators Olsson [2005]

Uncertainty and imprecise information: because environmental data collection is often expensive and difficult, measurement error is often large, and spatial and temporal sampling may not fully capture system behaviour. Records may also contain missing

values and highly uncertain information. See Section 3.1

Many environmental systems involve processes which are not yet well known, and for which no formal models are established at present. Because the consequences of an environmental system changing behavior or operating under abnormal conditions may be severe, there is a great need for Knowledge Discovery in the area.

Great quantities of data are available, but as the effort required to analyse the large masses of data generated by environmental systems is large, much of it is not examined in depth and the information content remains unexploited. The special features of environmental processes demand a new paradigm to improve analysis and consequently management. Approaches beyond straightforward application of conventional classical techniques are needed to meet the challenge of environmental system investigation. Data mining techniques provide efficient tools to extract useful information from large databases, and are equipped to identify and capture the key parameters controlling these complex systems.

3 DATA MINING TECHNIQUES

Here we shall introduce a variety of data mining techniques: classification (Section 3.5), clustering (Section 3.4), and association rule extraction (Section 3.6), as well as preprocessing other data issues. Of course, we cannot hope to detail all data mining tools in a short paper. An extensive review of data mining tools for environmental science is given in Spate and Jakeman [Under review 2006], and references to specific papers are given throughout the text. Key reading material introducing the reader to essential points of KDD are Han and Kamber [2001], Whitten and Frank [1991], Hastie et al. [2001], Larose [2004] and Parr Rud [2001]. The techniques listed below are some of the most common and useful in the data mining toolbox, and preprocessing and visualisation are also included in this section, as they are essential components of the knowledge discovery process.

3.1 Preprocessing: Data Cleaning, Outlier Detection, Missing Value Treatment, Transformation and Creation of Variables

Sometimes, a number of cells are missing from the data matrix. These cells may be marked as a *, ?, NaN (Not a Number), blank space or other special character or special numeric code such as 99999. The latter can produce grave mistakes in calculations if not properly treated. It is also important to distinguish between random and non-random missing values (Allison [2002], Little and Rubin [1987]). Non-random missing values are produced by identifiable causes that will determine the proper treatment, also influenced by the goals of the task. Inputation (see Rubin [1987]) is a complex process for converting missing data into useful data using estimation techniques. It is important to aviod false assumptions when considering inputation methods, because this choice may have a significant effect on the results extracted.

Options include removing the object altogether (although useful data may be thrown away), replacing it with a mean or otherwise estimated value, duplicating the row once for each possible value if the variable is discrete, or excluding it from the analysis by modification of the data mining algorithm. None of these are without pros and cons, and the choice of method must be made with care. In particular, removing rows with missing cells from a dataset may cause serious problems if the missing values are not randomly distributed. Caution should be exercised when adopting this approach, and it is of utmost importance to report any elimination performed.

Outliers are objects with very extreme values in one or more variables (Barnett and Lewis [1978]). Graphical techniques were once the most common method for identifying them, but increases in database sizes and dimensions have led to a variety of automated techniques. The use of standard deviations is possible when and only when considering a single variable that has a symmetric distribution, but outliers may also take the form of unusual combinations of two or more variables. The data point should be analysed as a whole to understand the nature of the outlier.

Outliers can then be considered apart treated as missing, corrected (this may be possible where decimal points have been mistakenly entered, for example), or included in the study as regular points. Whichever course of action is taken, the presence of an outlier should be noted for future reference. Certain modelling methods and data mining algorithms may be affected to a greater or lesser degree by the presence of outliers, a concern which should feature the choice of tools used throughout the rest of the process. See Moore and McCabe [1993] for an interesting discussion on the dangers of eliminating rows with outliers:

"in 1985 British scientists reported a hole in the ozone layer of the Earth's atmosphere over the South Pole. [...] The British report was at first disregarded, since it was based on ground instruments looking up. More comprehensive observations from satellite instruments looking down had shown nothing unusual. Then, examination of the satellite data revealed that the South Pole ozone readings were so low that the computer software [...] had automatically suppressed these values as erroneous outliers! Readings dating back to 1979 were reanalyzed and showed a large and growing hole in the ozone layer [...] suppressing an outlier without investigating it can keep valuable out of sight."

Moore and McCabe [1993]

Sometimes, transformation of variables may assist analysis. For example, normality may be forced when using ANOVA, or, for ease of interpretation, variables with a large number of categorical labels can be grouped according to expert knowledge. Under some circumstances, discretisation of continuous variables is appropriate (eg *Age* into *Child* under 18 years, *Adult* between 18 and 65 years, *Elderly* over 65 years). Noise is often a critical issue, and especially with environmental data some bias may exist that can be removed with a filter. Transformations should always be justified and documented, and the biases that may be introduced noted (Gibert and Sonicki [1999]).

Creation of additional variables is also used to facilitate the knowledge discovery process under some circumstances. For example, see the decision tree of Figure 2. The object here is to determine whether or not a given year is or is not a leap year, and to do this, an efficient tree is built using the variables *YearModulo4* and *YearModulo100*. Where extra features are constructed in this way, expert knowledge is usually the guide. Exploratory variable creation without such assistance is almost always prohibitively time consuming, and as noted in Section 5, may obfuscate physical interpretation and exacerbate noise. Efficient techniques for data reduction, however, do exist and are well used.

3.2 Data Reduction and Projection

When the number of variables is too high to deal with in a reasonable way, which is not unusual in data mining context, it may be convenient to apply a data reduction method. This kind of technique consists of finding some set with the minimum number of variables that captures the information contained in the original data set.

This may be accomplished by eliminating some variables wholesale, or projecting the feature space of the original problem into a reduced fictitious space, with fewer dimensions. Principal Components Analysis (PCA) (see for example Dillon and Goldstein [1984]) is one of the best known techniques used for the latter purpose. Each principal component is a linear combination of the original variables, and the aim is to work with a reduced set of these, such that the loss of information is not great. It is important to note that interpretation of the new variables may be lost.

Regarding the former method, datasets may contain irrelevant or redundant variables. For example, in the daily weather dataset discussed in Spate et al. [2003], four temperature variables were recorded: maximum, minimum, mean, and grass. In the context of that study, all four contained much of the same information and any three out of the four could be eliminated without significant loss of predictive capacity. A Boolean (presence/absence) marker for frost was also included in the original database, but this was only relevant for three out of six sites, as the other measurement stations where located in a subtropical maritime environment (the Brisbane region of Queensland, Australia) where frosts do not occur. No useful information was encoded in the Brisbane frost variables, which were all a vector of zeros and could thus be deleted from the Brisbane datasets.

Automated techniques for identifying and removing unhelpful, redundant or even contradictory variables usually take one of two forms: statistical examination of the relevance properties of candidate variables and combinations of the same, or searching through the space of possible combinations of attributes and evaluating the performance of some model building algorithm for each combination. The former are called *filters* and the latter *wrappers* (see Hall [1999] for details). For a survey of common attribute selection techniques, see Molina et al. [2002].

Other techniques are based on feature weighting

Aha [1998] Nez et al. [2003], which is a more general and flexible approach than feature selection. The aim of feature weighting is to assign a degree of relevance, commonly known as a weight, to each attribute. This way, some similarity computations for tasks like clustering, rule induction, can be improved. Similarities (or dissimilarities) become emphasized according to the relevance of the attribute, and irrelevant attributes will not influence the results, so quality of inductive learning improves.

3.3 Visualisation

While automation is a key goal for knowledge discovery routines, some human interaction is still beneficial and indeed necessary. One of the key points at which human interaction is often most fruitful is the visualisation stages, during pre- and postprocessing. Graphical methods should be the first stage of investigation for all datasets, even those whose dimension is too great to allow a comprehensive survey in this way. The presence of outliers, missing values, errors, and unusual behaviour are often first noted visually, enabling more detailed investigation later. Redundant and useless variables may also become clear at this stage, although by no means is visualisation a complete substitute for quantitative exploratory analysis.

Graphs commonly used for classical exploratory visualisation like boxplots, histograms, time series plots, and two dimensional scatter plots may be useful for examining individual variables or pairs of variables, but when considering a great number of variables with complex interrelations other devices may have greater utility, as scatter plots can usefully contain only a small number of dimensions and a limited number of points. A variety of more sophisticated visualisation methods appear in the context of data mining, for example:

- Distributional plots
- Three, four, and five dimensional plots (colour and symbols may be used to represent the higher dimensions)
- Using transformed variables, for example log scales
- Rotatable frames
- Animation with time

Many data mining packages (for example Weka of Waikato [2005]) include visualisation packages, and the more complex devices mentioned above such as rotatable reference frames for three dimensional plots and animations, can be generated with common packages such as Matlab (**?**]) or a dedicated data language such as IDL or the CommonGIS tool (Andrienko and Andrienko [2004]). There are also dedicated visualisation tools such as XGobi (Swayne et al. [1998]).

Visual representations are extremely effective, and may convey knowledge far better than numerical information or equations. As it is well accepted that presentation of the results from almost all modelling processes should include graphical illustrations, and we argue that the same approach is equally essential to the knowledge discovery process.

3.4 Clustering and Density Estimation

Clustering techniques are used to divide a data set into groups. They are suitable for discovering the underlying structure of the target domain, if this is unknown. For this reason, they belong to the group of techniques known as *unsupervised learners* along with association rule extraction, which will be discussed in Section 3.6. Clustering techniques cover an exploratory goal, rather than a predictive one. They identify distinct groups of similar objects (according to some criteria) that can be considered together, which is very useful in the Data Mining context, since the number of cases to be analysed can be huge. Ideally, objects within a cluster should be homogeneous compared to the difference between cluster representatives.

The measure of distance or dissimilarity between data objects can be based either a quantitative metric, dissimilarity measure, or some logical criteria derived from analogy or concept generalization, depending on the research field where the clustering algorithm was conceived (usually either Statistics (Sokal and Sneath [1963]) or Artificial Intelligence (Michalski and Stepp [1983])). Sometimes it is convenient to mix algebraic and logical criteria for better capturing difficult domain structures (Gibert et al. [2005a]). Note that where data is continuous and the scales changes between variables, normalisation may be necessary to avoid weighting variables unevenly. Appropriate choice of criteria for comparing objects (distance measure) is essential, and different measures will result in different clustering schemes, a point which is discussed in detail in Spate [In preparation], Núñez et al. [2004], and Gibert et al. [2005b].

There are different families of techniques, to be used depending on the desired form of the clusters. The simplest methods simply divide the feature space into a set number of (hyper) polygonal partitions, but other methods can construct overlapping or fuzzy classes or a hierarchy of clusters. For a survey, see Dubes and Jain [1988].

Clustering can also be viewed as a density estimation problem by assuming that the data was generated by a mixture of probability distributions, one for each cluster (see, e.g., Witten and Frank, 2005). A standard approach is to assume that the data within each cluster are normally distributed. In that case a mixture of normal distributions is used. To get an improved and more concise description of the data other distributions can be substituted when the assumption of normality is incorrect. The overall density function for the data is given by the sum of the density functions for the mixture components, weighted by the size of each component. The beauty of this approach is that one can apply the standard maximum likelihood method to find the most likely mixture model based on the data. The parameters of the maximum likelihood model (for example, means and variances of the normal densities) can be found using the expectation maximization algorithm. Treating clustering as a density estimation problem makes it possible to objectively evaluate the model's goodness of fit, for example, by computing the probability of a separate test set based on the mixture model inferred from the training data. This approach also makes it possible to automatically select the most appropriate number of clusters.

3.5 Classification and Regression Methods

In classification and regression, the identity of the target class is known *a priori* and the goal is to find those variables that best explain the value of this target, either for descriptive purposes (better understanding the nature of the system) or prediction of the class value of a new datapoint. They are an example of *supervised* learning methods. A popular and accessible classification model is the decision tree, of which an example is given in Figure 2. This simple model, built on two variables, tells us if a given year is a leap year or not. The information encoded is thus: if the year is divisible by four with no remainder (YearModulo4 = 0), that year is a leap year unless it is also exactly divisible by 100 (YearModulo100 = 0). The decision is given at the internal nodes of the tree, and if the condition holds, we follow the right-hand branch. If it fails, we follow the left-hand branch. In this way, the tree splits the dataset into smaller and smaller pieces until each can be assigned a label, which is the value of the target variable. Decisions can take the form of greater than/less than criteria, or equality with a specific value/s or category/ies. Perhaps the most common decision tree extraction algorithm is C4.5 (Quinlan [1993]).

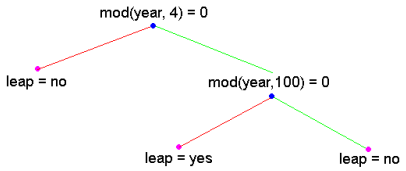

Figure 2: Example Decision Tree

Classical linear regression is a technique for finding the best linear equation defining the relationship between a numerical response variable and the independent variables, all of which should also be numerical (Draper and Smith [1998]). It is mainly used for prediction of the target variable, but also for identifying which variables have the strongest influence on the behavior of the response variable. In that sense it is useful for descriptive purposes. However, there are many conditions to be met in order for regression to be a suitable technique, such as normality, homocedasticity or independence of the regressors. When all the independent variables are qualitative, ANOVA should be used, and where both qualitative and numerical data are involved, ANCOVA is the proper model, provided the required technical hypotheses hold and the qualitative variables are properly transformed into dummy variables. For other situations, nonlinear regression may be useful (if nonlinearity is not polynomial, neural networks may be a better approach). For detailed discussion of multivariate regression, see for example Lebart et al. [1984] and Dillon and Goldstein [1984]. When modelling non numerical responses, it is possible to use logistic regression (when the response is binary or can be transformed to binary) or even polytomous regression (for qualitative responses of more than 2 categories). In that case, interpretation of results requires significant care.

A middle way between the two approaches of classification and regression exists in the family of methods known as regression trees. Here, the dataset is split up into blocks by a tree like the one shown in

Figure 2, but instead of a class label on each leaf, there is a model obtained by regression. The M5' (em five prime) mentioned in Frank et al. [2000] is an example of this concept.

Case-Based Reasoning (CBR) is a general problem-solving, reasoning and learning paradigm (see Kolodner [1993]) within the artificial intelligence field. CBR relies on the hypothesis that similar problems have similar solutions, and new problems can be solved from past solutions stored in an experience memory, usually called the case library or case base. If the structure of the cases or experiences is 'flat', where a case can be described with a set of pairs formed by one attribute and its value, the application of CBR principles is generally known as the nearest neighbour technique, memory-based reasoning or instance-based reasoning. CBR can be used as a classification method. In this case, the assumption is that similar cases should have similar classifications: given a new case, one or more similar cases are selected from in the case library, and the new case is classified according to the classifier values of those neighbours. Quality of the case library is critical for good classification, as is the choice of an appropriate measure of similarity between cases (see Núñez et al. [2004]). In CBR systems, the retrieval of similar instances from memory is a crucial point. If it is not carried out with accuracy and reliability, the system will fail. The retrieval task is strongly dependent on the case library organization (Sànchez Marrè et al. [2000]). It must also be noted that CBR does not produce an explicit model describing system behaviour in the way that most models do. Rather, the model is implicit in the database itself.

Standard feed-forward neural nets are supervised learners, although another form of neural net (the self-organising map) exists as an unsupervised method. They can be used for classification and for regression. Most simply, neural nets consist of a series of nodes arranged in a number of layers, commonly three- an input layer, a single hidden layer where nonlinear transformations of the input variables are performed, and an output. Additional layers of hidden nodes and feedback can be introduced into the system, providing greater flexibility at the cost of increased complexity. This complexity is the chief reason for the idea that neural nets are 'black boxes', which is not entirely warranted. With good visualisation, process information can be extracted from them.

As noted in the introduction, artificial neural networks (also called neural nets) have been extensively applied in the environmental sciences, and for that reason we will not examine them in detail here. Numerous references can be found in Section 4.

Rule induction or rule extraction is the process of discovering rules summarising common or frequent trends within a dataset (i.e. which variables and values are frequently associated). Classification rules are induced rules from labelled examples. The examples should be marked with the cluster or the class label. Thus, it is a supervised machine learning technique, which can be used to predict the cluster to which a new example or instance belongs to. The induced rules are not usually guaranteed to cover the entire dataset, and focus more on representing trends in the data. Classification rule extraction algorithms are similar to association rule discovery techniques discussed in Section 5. In an environmental context, supervised classification rules could be used, for example, to identify from a grid of spatial points those locations that may be prone to gully erosion. A rule for erosion vulnerability might look like:

IF slope $>$ 10 % AND soiltype = yellow podzolic AND landuse = grazing THEN risk = high

Note that some rule extraction routines can combine numerical and categorical data. A time component can also be introduced into the rule format. Consider the case where, for a catchment of a given area A, the time delay between rainfall P and streamflow Q events is T, or in other words, IF P AND A THEN Q WITHIN TIME T. As a formulaic expression, this rule might be compactly expressed as:

$$P, A \Rightarrow Q^T$$

3.6 Association Analysis

Association analysis is the process of discovering and processing interesting relations from a dataset. The concept was originally developed for supermarket analysis, where the aim is to uncover which items are frequently bought together. They have the advantage that as no initial structure is specified, the results may contain rules that are highly unexpected and which would never have been specifically searched for because they are inherently surprising. The format summarising only frequently occurring patterns can also be useful for anomaly detection, because those datapoints violating rules that usually hold are easy to identify and may be examples of interesting behaviour.

Rule extraction algorithms, for both association and classification, tend to fall into two broad categories: those built by generalising very specific rules until they cover a certain number of instances (for example the AQ family of algorithms described in Wnek and Michalski [1991], and those that begin with a broad rule covering all or a large fraction of the data and refine that rule until a sufficient level of precision is achieved (such as the PRISM Cendrowska [1998] and RIPPER Cohen [1995] algorithms. For obvious reasons, the specific to general variety are for the most classification rule learners. Many rule extraction algorithms are extremely fast, and can thus be applied to very large databases in their entirety. They may be used either for predictive purposes or for system investigation.

3.7 Spatial and Temporal Aspects of Environmental Data Mining

The dominating field of application for the knowledge discovery process is that of business. Only recently has data mining expanded into other fields, as the utility of knowledge extraction methods has been noted by researchers outside the data mining community. In business, knowledge extraction techniques are used for identifying consumer patterns or maximizing profit-related functions. Data is typically sourced from business transactions databases or Enterprise Resource Planning (ERP) systems. Spatial and temporal considerations are usually not significantly related to business goals, although they may be included during the data filtering/cleaning preparation phases (i.e. with a marker for transactions recorded in July, or in Vermont). This kind of spatial data can be easily integrated into a dataset for knowledge discovery, as can time or distance to a specific point. However, environmental science often requires deep consideration of spatial and temporal variables, and indeed this is often the primary focus. Therefore, more sophisticated spatiotemporal techniques are valuable.

Temporal relationships include ”before-during-after” relations, while spatial relationships deal either with metric (distance) relations, or with non metric topological relations. Classical data mining methods do not accommodate this kind of need, but many can be modified to accept it (and have been, see for example Antunes and Oliveira [2001], Spate [2005]) with minimal effort and good results. Spatial data mining is often treated as an extension of temporal data mining techniques (including time series analysis and sequence mining) into a multidimensional space.

Recently, data mining techniques have been designed with spatial and temporal data in mind, and a significant body of spatial and temporal data mining research does exist. Time series data mining techniques have been built for stock market data extraction and industrial purposes (for example see Antunes and Oliveira [2001], Chen et al. [2004]), and even environmental science (see Sànchez Marrè et al. [2005]). It has been stated in the literature (Keogh and Kasetty [2002]) that temporal data mining is an area where much effort should be focussed. An example of an explicitly spatial machine learning technique is Cellular Automata, with their interacting grid points where spatial and temporal dynamics can be modelled very naturally. For an example application, see the bird nest site work of Campbell et al. [2004]. Here, consideration of neighbouring nest sites was an integral part of the model. Some other spatial and temporal data mining examples are listed in Section 4.

Spatiotemporal data mining is potentially useful for a variety of tasks, including: **a** spatiotemporal pattern identification (as in pattern analysis, neighborhood analysis) **b** data segmentation and clustering (spatiotemporal classification) **c** dependency analysis, correlation analysis and fault detection in data (outlier detection, surprising pattern identification) **d** trend discovery, sequence mining (as in regression analysis and time series prediction). It should be noted that spatial resolution and time granularity affects the nature of the extracted patterns and must be chosen with appropriate care. In this respect, visualizing data and extracted patterns, employing maps and GIS technology (from example see Andrienko and Andrienko [2004]) could be valuable.

4 PREVIOUS WORK IN THE AREA

As mentioned in the introduction, most data mining techniques have not found wide scale application in the environmental sciences. In this section we mention a few projects that have utilised this technology, although as with all literature reviews, we do not claim to make an exhaustive list. A few papers have been published advocating learning methods for environmental applications, for example Babovic [2005]. Comas et al. [2001] discusses the performance of several data mining techniques (decision tree creation, two types of rule induction, and instance-based learning) to identify patterns from environmental data. A small number of research groups also exist with the specific aim of using artificial intelligence or data mining in the environmental sciences.

The BESAI (Binding Environmental Science and Artificial Intelligence) working group has organised four international workshops within the ECAI conferences during 1998-2004, one international workshop at IJCAI'2003 conference, and one international workshop at AAAI'99, with contributions addressing data mining techniques. They have also organised two special sessions devoted to Environmental Sciences and Artificial Intelligence during the iEMSs 2002 and iEMSs 2004 international conferences. See the BESAI website for more details (BESAI, http://www.lsi.upc.edu/ webia/besai/besai.html)

The European Network of Excellence on Knowledge Discovery (KDnet) organised a workshop on Knowledge Discovery for Environmental Management Voss et al. [2004] in an effort to promote KDD in the public sector. Four International Workshops on Environmental Applications have been held over the period 1997 – 2004, producing some of the papers discussed below. The output of the latest workshop, which focussed on genetic algorithms and neural networks, is discussed in Recknagel [2001].

One exception from the statement that data mining techniques are not widely used in the area is the use of Artificial Neural Networks, which have become an accepted part of the environmental modelling toolbox. For examples see Kralisch et al. [2001] and Almasri and Kaluarachchi [2005] on nitrogen loading, Mas et al. [2004] on deforestation, or Belanche et al. [2001], Gibbs et al. [2003] and Gatts et al. [2005] on water quality, or the discussion in Recknagel [2001]). Numerous other examples can be found in most journals in the area.

Examples of the use of clustering algorithms include Sànchez-Marrè et al. [1997], where different techniques for clustering wastewater treatment data were compared, Zoppou et al. [2002], where 286 Australian streamflow series were clustered according to a dimensionally reduced representation. The aim was to identify groups of catchments with similar physical characteristics. Clustering or similar methods have also been used to a similar end in and [????] and Sanborn and Bledsoe [2005] for study areas in the United Kingdom and Northwestern United States respectively. Clustering was also applied to cyclone paths in Camargo et al. [2004], and in Ter Braak et al. [2003] to cluster water samples according to chemical composition.

Various classification algorithms have been applied to a wide variety of environmental problems as well. Rainfall intensity information was extracted from daily climate data Spate [2002] and Spate et al. [2003] using a number of classification methods. A decision tree like that in Figure 2 was used in Ekasingh et al. [2003] where the cropping choices of Thai farmers were modelled as a classification problem with considerable success. Mosquito population sites were classified in Sweeney et al. [2004 (submitted], with a view to controlling the spread of malaria. Agriculturally, classification has been applied to apple bruising (Holmes et al. [1998]), mushroom grading (Cunningham and Holmes [1999]), bull castration and venison carcase analysis in Yeates and Thomson [1996], and perhaps most famously in Michalski and Chilausky's classic soybean disease diagnosis work (Michalski and Chilausky [1980]). Regression trees were applied (for example) to sea cucumber habitat preference modelling in Dzeroski and Drumm [2003].

Much effort has been made by international scientists in the water quality and wastewater quality control domains. Some approaches used rule-based reasoning (Zhu and Simpson [1996]), case-based reasoning (Rodríguez-Roda et al. [1999]), fuzzy logic (Wang et al. [1997]), artificial neural networks (Syu and Chen [1998]), and integrated approaches were also developed, such as in (Rodríguez-Roda et al. [2002]) and (Cortés et al. [2002]). Many of these approaches utilised several data mining techniques.

In the study of urban air quality, fuzzy lattice classifiers have been applied for estimating ambient ozone concentrations in an operational context, with very good results Athanasiadis et al. [2003]. Uncertainty and other data quality issues such as measurement validation and estimation of missing values in the same field were addressed in Athanasiadis and Mitkas [2004]. A comparison between statistical and classification algorithm algorithms applied in air quality forecasting Athanasiadis et al. [2005] demonstrated that the potential of data mining techniques is high.

Classification has also found spatial applications. For example, fish distribution (Su et al. [2004]) and soil erosion patterns (Ellis [1996]) have both been modelled with classification methods, as was soil erosion in Smith and Spate [2005], and other soil properties in McKenzie and Ryan [1999], which also used regression trees and other techniques with a view to obtaining system information.

The use of rule learning for the environmental sciences is discussed in Riaño [1998]. The example discussed in this paper is the state identification of a

wastewater treatment plant. Rule learning was also used to investigate a streamflow/electrical conductivity system in Spate [2005]. In Dzeroski et al. [1997], the process of rule learning is illustrated with examples from water quality databases and the CN2 algorithm. Rodríguez-Roda et al. [2001] presented the induction of rules in order to acquire specific knowledge from (bio)chemical processes).

Data mining and machine learning are of course not restricted to the methods discussed here, and some less common techniques have been applied to environmental problems. In Robertson et al. [2003], Hidden Markov models were used to model rainfall patterns over Brazil with interesting results, and Mora-López and Conejo [1998] applied qualitative reasoning to meteorological problems. Cloud screening for meteorological purposes was also investigated with Markov Random Fields in Cadez and Smyth [1999]. Sudden death of oak trees was modelled with support vector machines in Guo et al. [2005]. Generative topographic mapping was used to investigate riverine ecology in Vellido et al. [submitted 2005]. Genetic programming was used to model glider possum distributions Whigham [2000]), and D'heygere et al. [2003] used genetic algorithms for attribute selection in benthic macroinvertebrate modelling. Decision trees were then built from the reduced dataset. Several inductive methods have been applied to discover knowledge of the behaviour of wastewater treatment plants, such as in Comas et al. [2001].

5 GOOD DATA MINING PRACTICE

As with all modelling paradigms, good practice modelling involves far more than applying a single algorithm or technique. Each of the steps detailed in Section 1 must be followed with due attention. In this section, we record a few notes and considerations that may be of use to those contemplating the use of data mining in an environmental area.

Data Cleaning Data cleaning is a fundamental aspect of the analysis of a dataset, and one which is often neglected. When working with real data, the process is often very time consuming, but is essential for obtaining good quality results, and from there useful new knowledge. The quality of the results directly depends on the quality of the data, and in consequence, on the correct missing data treatment, outlier identification, etc. Data miners should become conscious of the importance of performing very careful and rigorous data cleaning, and allocate sufficient time to this activity accordingly.

Transformations Beginning with preprocessing, avoidance of unnecessary transformations is recommended, especially if the transformation decreases interpretability (for example $Y = \log streamflow$, although Y is normal). If transformations are definitely required, some bias may be introduced into the results; thus, it is convenient to minimize arbitrariness of the transformation as much as possible (in recoding *Age*, Adult may be defined from 18 to 65 or from 15 to 70), and this implies that the goals of the analysis must also be taken into account. For arithmetic transformations, inputation of missing data before the transformation is thought to be better, especially if several variables are combined into one new feature (a mean or ratio, from example).

The same caveat applies to interpolated or extrapolated data, for example to catchment-wide rainfall estimates obtained by thin plate spline interpolation. It is not uncommon for no mention at all to be made of the fact that rainfall values used are not directly obtained, despite the possibility of bias and error being introduced by the interpolation process.

Input Data Uncertainty All data is subject to uncertainty, and environmental data such as rainfall or streamflow are often subject to uncertainties of $\pm$ 10% or more. Tracking and reporting of uncertainties related to measurement and other sources of noise is another area that is sometimes not treated rigorously, despite the implications. Consider the following: there is a $\pm$ 10% error in all input data. Therefore, the minimum theoretically achievable error of any model built on this data cannot be less than $\pm$ 10%, and is likely to be much higher depending on the structure of the model. Models with reported fit greater than this are overfitted and their performance measures do not reflect true predictive capacity.

Quantity of data is also a concern. In general, where there is more data, there is less uncertainty, or at least that uncertainty can be better quantified. Choice of data mining method should also be influenced by dataset size. Where datasets are small, choose simpler methods and be mindful of the maximum theoretical certainty that can be obtained. It is also important to remark that as the number of data increases, variance of classical estimators tends to zero, which usually implies that very small sample differences may appear statistically significant. This phenomena requires serious attention and great care must be exercised in the interpretation of some statistical results, and it requires the user to take into account that statistical significance properly reflect

the nature of the data. In fact, serious revision of classical statistical inference is necessary to enable suitable use in the context of data mining.

Data Reduction by Principal Components and Similar Techniques Principal component analysis is only recommended when all original variables are numerical. For qualitative data, multiple correspondence analysis should be used in its place. See for example Lebart et al. [1984] or Dillon and Goldstein [1984] or methodological details. This process reduces the original variables to a set of fictitious ones (or factors). With principal component analysis, conceptual interpretation of a factor may not be clear, and if this is the case, there will be implications for understandability of the final results. Numerous other techniques also exist for feature weighting and and selection.

Clustering Most clustering methods generate a set of clusters even where no set of distinguishable groups really exist. This is why it is very important to carefully validate the correctness of the discovered clusters. Meaning and usefulness of discovered classes are one validation criteria, although this is largely subjective. A more quantitative approach is to perform multiple runs of the algorithm or different algorithms with slightly different parameters or initial values, which will give a good indication of the stability of the cluster scheme. Some software packages also contain tools to help assess such properties.

As a measure of cluster 'goodness', the ratio of average distance within to average distance between clusters may be useful where a numerical distance measure exists, although it is redundant if that criteria was used to build the clusters themselves (as is the case of Ward's method Ward [1963]). Cluster validation where no reference partition exists (and in real-world applications none is present, or the clustering would be unnecessary) is an open problem, but some investigation into stability should be performed as a minimum treatment. See for example Gibert et al. [2005c]. Some methods based on the principles of cross validation can also be used to analyze how the representatives of the classes move from one iteration to another. See Spate [In preparation] for an example of this procedure.

Statistical Modelling and Regression Scalar real-valued performance criteria such as the *determination coefficient* (R^2, also known as efficiency), used together with residual plots (Moore and McCabe [1993]), constitute a very useful tool for validation of the model, far more powerful than numerical indicators by themselves. Outliers, influent values, nonlinearities and other anomalies can be investigated in this way. Note however that R^2 can be applied only to real numerical data.

Classification When classifying real data, it is often useful to consider accuracy on a class-by-class basis. In this way, the modeller can keep track of where errors are occurring. These errors may be given unequal weighting, if the consequences are not equal. The most common device for this is the confusion matrix. If the problem contains only two classes (say true/false), the matrix is filled with the following entries:

Top Left 'true' values correctly labelled 'true'

Top Right 'true' values incorrectly labelled 'false'

Bottom Left 'false' values correctly labelled 'false'

Bottom Right 'false' values incorrectly labelled 'true'

The distribution of input data should also receive consideration, as many classification algorithms tend towards predicting the majority class. An in-depth discussion of this topic can be found in Weiss and Provost [2001]. Tree (and other classifier) stability can be assessed in the same ways as cluster stability (see above).

Uncertainty Quantification and Model Validation As mentioned in the note regarding input data above, proper consideration of uncertainty is essential for meaningful modelling. One must also give thought to how best to quantify and represent the performance of the final model. For some purposes, a single-valued measure such as R^2 may be sufficient provided that the model has been properly validated as unbiased, but for most applications more information is useful. It is seldom possible to represent model performance against all goals of the investigation with one number. As an example, a rainfall-runoff model may fit well for low and medium flows, but underestimate large peaks. It may also have a systematic tendency to slightly overpredict lower flows to compensate for missing extreme events. All of this cannot be expressed as a single number, but a comparison of distributions will reveal the necessary information.

Model validation is as important for automatically extracted models as it is for those constructed with more human interaction, or more so. To this end we

recommend the usual best practice procedures such as holding back a portion of the dataset for independent validation (if the size of database allows) and n-fold cross validation.

Parameter Selection and Model Fitting While parameter-free data mining algorithms do exist, most require some *a priori* set up. Parameters for data mining algorithms are decided by the same methods as more common models- expert knowledge, guessing, trial and error, automated and manual experimentation. In addition, it is often helpful to learn a little about the role of the parameter within the algorithm, as appropriate values for the problem at hand can often be set or estimated this way. Some experimentation may improve the output model and reporting the process of parameter fitting in detail adds credibility to any modelling project. It is important that parameter values are not chosen based on the final test data. Otherwise optimistic performance estimates will be obtained.

6 Challenges for Data Mining in the Environmental Sciences

Finally, in this section, we shall comment on the hot issues and challenging aspects in and of the interdisciplinary field of environmental data mining sciences in coming years. Achievement of the following aims would increase utility and applicability of data mining methods.

- Improvement of automated preprocessing techniques
- Elaboration of protocols to facilitate sharing and reuse of data
- Development of standard procedures (benchmarks) for experimental testing and validation of data mining tools
- Involvement of end-user (domain expert) criteria in algorithm design and result interpretation
- Development and implementation of mixed data mining methods, combining different techniques for better knowledge discovery
- Formulation of tools for explicit representation and handling of discovered knowledge for greater understandability
- Improvement of data mining techniques for on-line and heterogenous databases
- Guideline or recommendation development, to assist with method and algorithm selection.

Another factor that is often of great importance is (conceptual) interpretability of output models. Cutting edge knowledge acquisition techniques such as random forests have advantages over simpler methods, but are difficult to interpret and understand. Tools that clearly and usefully summarise extracted knowledge are of great value to environmental scientists, as are those that assist in the quantification of uncertainties.

6.1 Integrated Approaches

The main goal of many environmental system analyses is to support posterior decision making to improve either management or control of the system. Intelligent Environmental Decision Support Systems (IEDSSs) are among the most promising approaches in this field. IEDSS are integrated models that provide domain information by means of analytical decision models, and allow access to databases and knowledge bases to the decision maker. They intend to reduce the time in which decisions can be made as well as repeatability and the quality of eventual decisions by offering criteria for the evaluation of alternatives or for justifying decisions Poch et al. [2004a], Poch et al. [2004b]. Often, multiple scenarios are modelled and evaluated according to environmental, social, and economic criteria.

There are six primary approaches to the problem of building an integrated model: expert systems, agent-based modelling, system dynamics, Bayesian networks, coupled complex models, and meta-modelling. Of these, the last three are most relevant to the field of data mining. Opportunities exist for automation of Bayesian network and meta-model construction and parametrisation, simplification and summarisation of complex submodels, and also interpretation of results. Data mining techniques are important tools for knowledge acquisition phase of of integrated model building, and because integrated models are very high in complexity, results are often correspondingly difficult to interpret and the decision maker may benefit from a postprocessing data mining step. Of course, data mined models may also form part of the integrated model as in Ekasingh et al. [2005].

7 SOFTWARE- EXISTING AND UNDER DEVELOPMENT

In this Section, two software packages are discussed in detail. One of these, Weka, is well established in the data mining community, and the other, GESCONDA, is currently under development. Weka is a general purpose package, and GESCONDA is designed specifically for environmental science. Both contain a wide variety of tools and techniques.

7.1 GESCONDA

GESCONDA (Gibert et al. [2004], Sànchez-Marrè et al. [2004]) is the name given to an Intelligent Data Analysis System developed with the aim of facilitating Knowledge Discovery (KD) and especially oriented to environmental databases. On the basis of previous experiences, it was designed as with four level architecture connecting the user with the environmental system or process. These four levels are the following:

- Data Filtering
  - data cleaning
  - missing data management
  - outlier analysis and treatment
  - statistical univariate analysis
  - statistical bivariate analysis
  - visualization tools
  - attribute or variable transformation facility
- Recommendation and Meta-Knowledge Management
  - overall goal definition
  - method suggestion
  - parameter setting
  - integration of attribute and variable metadata
  - domain theory and domain knowledge elicitation
- Knowledge Discovery
  - clustering (by machine learning and statistical means)
  - decision tree induction
  - classification rule induction

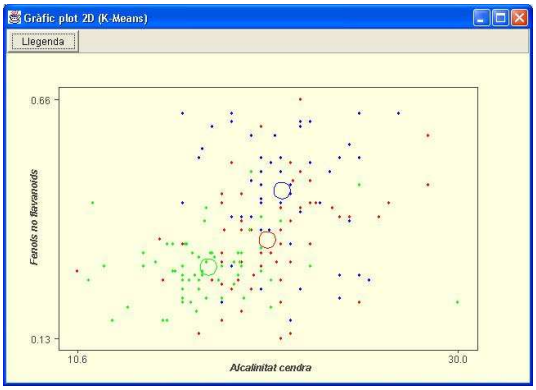

Figure 3: GESCONDA two-dimensional cluster plot

  - case-based reasoning
  - support vector machines
  - statistical modelling
  - dynamic systems analysis
- Knowledge Management
  - validation of the results
  - integration of different knowledge patterns for a predictive task, or planning, or system supervision, together with AI and statistics mixed techniques
  - consideration of knowledge use by end-users

Central characteristics of GESCONDA are the integration of statistical and AI methods into a single tool together with mixed techniques, for extracting knowledge contained in data, as well as tools for qualitative analysis of complex relationships along the time axis Sànchez-Marrè et al. [2004]. All techniques implemented in GESCONDA can share information among themselves to best co-operate for extracting knowledge. It also includes capability for explicit management of the results produced by the different methods.

Figure 3 is a two dimensional GESCONDA visualisation of a multidimensional clustering scheme and Figure 4 a screen capture from the clustering GUI. Figure 5 shows a rule induction screen. Portability of the software between platforms is provided by a common Java platform. The GESCONDA design document can be viewed at http://www.eu-lat.org/eenviron/Marre.pdf.

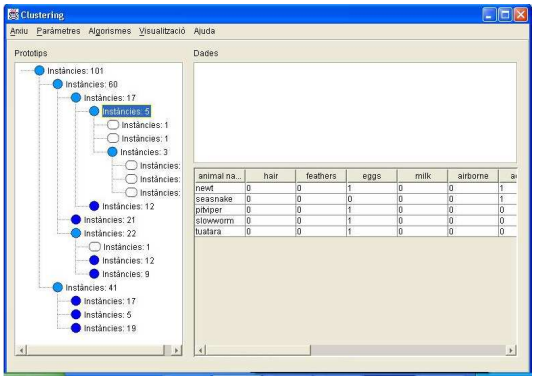

Figure 4: GESCONDA clustering interface

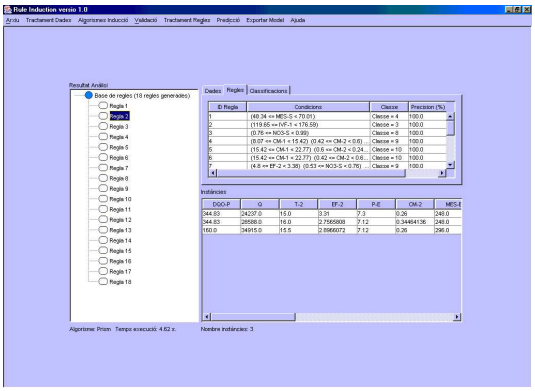

Figure 5: GESCONDA rule induction interface

7.2 Weka

The Weka workbench (Witten and Frank [2005]) contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality. A command line interface is also included, for mass processing. It was originally designed as a tool for analyzing data from agricultural domains but is now used in many different application areas, largely for educational purposes and research. The main strengths of Weka are that it is (a) freely available under the GNU General Public License, (b) very portable because it is fully implemented in the Java programming language and thus runs on almost any computing platform, (c) contains a comprehensive collection of data preprocessing and modelling techniques, and (d) is easy to use by a novice due to the graphical user interfaces it contains.

Weka supports several standard data mining tasks. More specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection are included. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data object is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is a separate piece of software for converting a collection of linked database tables into a single table that is suitable for processing using Weka (Reutemann et al. [2004]). Another important area that is currently not covered by the algorithms included in the Weka distribution is time series modelling.

Weka's main user interface is the Explorer, shown in Figure 6, but essentially the same functionality can be accessed through the component-based Knowledge Flow interface, shown in Figure 7, and from the command line. There is also the Experimenter, which allows the systematic comparison of the predictive performance of Weka's machine learning algorithms on a collection of datasets rather than a single one.

The Explorer interface has several panels that give access to the main components of the workbench. The **Preprocess** panel has facilities for importing data from a database as a CSV or other format file, and for preprocessing this data using a so-called **fil-**

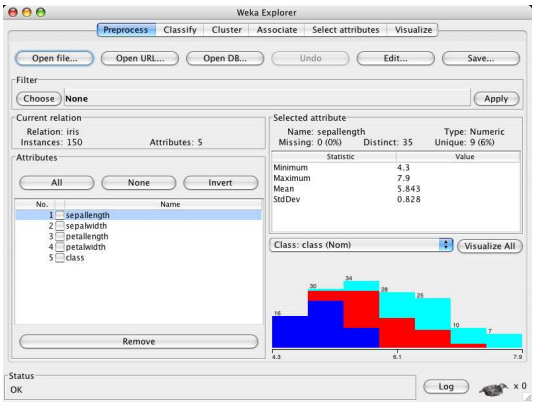

Figure 6: The Weka Explorer user interface

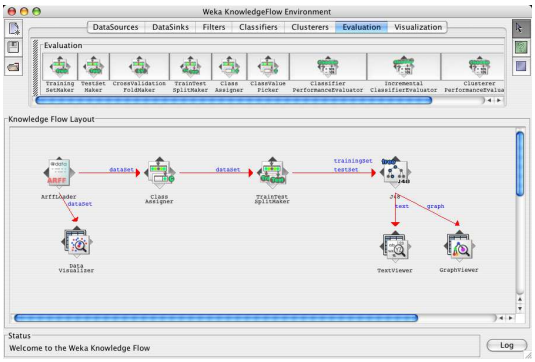

Figure 7: The Weka Knowledge Flow user interface

tering algorithm. These filters can be used to transform the data (e.g. turning numeric attributes into discrete ones) and make it possible to delete instances and attributes according to specific criteria. The **Classify** panel enables the user to apply classification and regression algorithms (indiscriminately called *classifiers* in Weka) to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, ROC curves, etc, or the model itself (if the model is amenable to visualization as, for example, decision trees are). The **Associate** panel provides access to association rule learners, which attempt to identify all important interrelationships between attributes in the data. The **Cluster** panel gives access to the clustering techniques in Weka, for example, the simple k-means algorithm. There is also an implementation of the expectation maximization algorithm for learning a mixture of normal distributions (see Section 3.4. The next panel, **Select attributes** houses algorithms for identifying the attributes in a dataset with most predictive capacity. The last panel, **Visualize**, shows a scatter plot matrix, where individual scatter plots can be selected and enlarged, and analyzed further using various selection operators.

7.3 Other

Other proprietary data mining packages exist. SAS's *Enterprize* miner GUI was designed for business users, and includes decision trees and neural nets within the wider SAS statistical framework and includes facility for direct connection with data warehouses. IBM has released *Intelligent Miner. Clementine* from SPSS includes facilities for neural networks, rule induction, data visualization in tables, histograms, plots and webs. Salford System's *CART* ® package builds classification and regression trees. Data mining libraries for general computation and statistics environments like Matlab and R have also been built, covering a range of techniques. These, with Weka and GESCONDA, are some of the data mining software options available.

8 AIMS OF IEMSS

iEMSs is the *International Environmental Modelling and Software Society* (www.iemss.org), founded in 2000 by interested scientists with the following aims:

- developing and using environmental modelling and software tools to advance the science and improve decision making with respect to resource and environmental issues;
- promoting contacts and interdisciplinary activities among physical, social and natural scientists, economists and software developers from different countries;
- improving the cooperation between scientists and decision makers/advisors on environmental matters;
- exchanging relevant information among scientific and educational organizations and private enterprises, as well as non-governmental organizations and governmental bodies.

To achieve these aims, the iEMSs:

- organizes international conferences, meetings and educational courses in environmental modelling and software;
- publishes scientific studies and popular scientific materials in the Environmental Modelling and Software journal (Elsevier);

- hosts a website (www.iemss.org) which allows members to communicate research and other information relevant to the Society's aims with one another and the broader community;
- delivers a regular newsletter to members.

This paper proposes that data mining techniques are valuable tools that could be used to good effect in the environmental and natural resource science field, and are thus of interest to iEMSs and its members. We aim to introduce the main concepts of data mining and foster discussion of the ways in which it could be used and encouraged within and outside the iEMSs organisation.

ACKNOWLEDGMENTS

The project $TIN2004 - 01368$ has partially financed the development of GESCONDA. Portions of Section 4 also appear in Spate and Jakeman [Under review 2006]. The aims and activities of iEMSs are taken with minimal alteration from the iEMSs website http://www.iemss.org/.

REFERENCES

Technical report, ????

Aha, D. Feature weighting for lazy learning algorithms. In Liu, H. and lastname Motoda, H., editors, *Feature Extraction, Construction and Selection: A Data Mining Perspective*. Kluwer, 1998.

Allison, P. *Missing Data*. Sage, Thousand Oaks, CA, USA, 2002.

Almasri, M. and J. Kaluarachchi. Modular neural networks to predict the nitrate distribution in ground water using the on-ground nitrogen loading and recharge data. *Environmental Modelling and Software*, 20(7):851–871, July 2005.

Andrienko, G. and A. Andrienko. Research on visual analysis of spatio-temporal data at fraunhofer ais: an overview of history and functionality of commonGIS. In *Proceedings of the Knowledge-Based Services for the Public Services Symposium, Workshop III: Knowledge Discovery for Environmental Managment*, pages 26–31. KDnet, June 2004.

Antunes, C. and A. Oliveira. Temporal data mining: An overview. In *KDD 2001: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2001. Workshop.

Athanasiadis, I., V. Kaburlasos, P. Mitkas, and V. Petridis. Applying machine learning techniques on air quality for real-time descision support. In *Information technologies in Environmental Engineering*, June 2003.

Athanasiadis, I., K. Karatzas, and P. Mitkas. Contemporary air quality forecasting methods: A comparative analysis between statistical methods and classification algorithms. In *Proceedings of the 5th International Conference on Urban Air Quality*, March 2005.

Athanasiadis, I. and P. Mitkas. Supporting the decision-making process in environmental monitoring systems with knowledge discovery techniques. In *Proceedings of the Knowledge-Based Services for the Public Services Symposium, Workshop III: Knowledge Discovery for Environmental Managment*, pages 1–12. KDnet, June 2004.

Babovic, V. Data mining in hydrology. *Hydrological Processes*, 19:1511–1515, 2005.

Barnett, V. and T. Lewis. *Outliers in Statistical Data*. Wiley, 1978.

Belanche, L., J. Valdés, J. Comas, I. Rodríguez-Roda, and M. Poch. Towards a model of input-output behaviour of wastewater treatment plants using soft computing techniques. *Environmental Modelling and Software*, 5(14):409–419, 2001.

Cadez, I. and P. Smyth. Modelling of inhomogeneous markov random fields with applications to cloud screening. Technical Report UCI-ICS 98-21, 1999.

Camargo, S., A. Robertson, S. Gaffney, and P. Smyth. Cluster analysis of western north pacific tropical cyclone tracks. In *Proceedings of the 26th Conference on Hurricanes and tropical Meteorology*, pages 250–251, May 2004.

Campbell, A., B. Pham, and Y.-C. Tian. Mining ecological data with cellular automata. In *Proceedings of the International Conference on Cellular Automata for Research and Industry (ACRI 2004)*, pages 474–483. Springer, October 2004.

Cendrowska, J. Prism: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies*, 27(4):349–370, 1998.

Chen, J., H. He, G. Williams, and H. Jin. Temporal sequence associations for rare events. In Honghua, D., namesleft>1 Srikant, R.,numnames>2 and lastname Zhang, C., editors, *Advances in Knowledge Discovery and*

Data Mining, 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004, Proceedings, volume 3056 of *Lecture Notes in Computer Science*, pages 239–239. Springer, 2004.

Cohen, W. Fast effective rule induction. In Prieditis, A. and lastname Russell, S., editors, *Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann, 1995.

Comas, J., S. Dzeroski, and K. Gibert. Knowledge discovery by means of inductive methods in wastewater treatment plant data. *AI Communications*, 14(1):45–62, 2001.

Cortés, U., I. Rodríguez-Roda, M. Sànchez-Marrè, J. Comas, C. Cortés, and M. Poch. DAI-DEPUR: An environmental decision support system for supervision of municipal waste water treatment plants. In *Proceedings of the 15th European Conference on Artificial Intelligence (ECAI'2002)*, pages 603–607, 2002.

Cunningham, S. and G. Holmes. Developing innovative applications in agriculture using data mining. In *Proceedings of the Southeast Asia Regional Computer Confederation Conference*, 1999.

D'heygere, T., P. Goethals, and N. De Pauw. Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates. *Ecological Modelling*, 160 (3):291–300, 2003.

Dillon, W. and M. Goldstein. *Multivariate Analysis*. Wiley, USA, 1984.

Draper, N. and H. Smith. *Applied Regression Analysis*. Wiley, 1998.

Dubes, R. and A. Jain. *Algorithms for Clustering Data*. Prentice Hall, 1988.

Dzeroski, S. and D. Drumm. Using regression trees to identify the habitat preference of the sea cucumber (holothuria leucospilota) on rarotonga, cook islands. *Ecological Modelling*, 170(2–3): 219–226, 2003.

Dzeroski, S., J. Grbovic, W. Walley, and B. Kompare. Using machine learning techniques in the construction of models. ii. data analysis with rule induction. *Ecological Modelling*, 95(1):95–111, 1997.

Ekasingh, B., K. Ngamsomsuke, R. Letcher, and J. Spate. A data mining approach to simulating land use decisions: Modelling farmer's crop choice from farm level data for integrated water resource management. In Singh, V. and lastname Yadava, R., editors, *Advances in Hydrology: Proceedings of the International Conference on Water and Environment*, pages 175–188, 2003.

Ekasingh, B., K. Ngamsomsuke, R. Letcher, and J. Spate. A data mining approach to simulating land use decisions: Modelling farmer's crop choice from farm level data for integrated water resource management. *Journal of Environmental Management*, 2005.

Ellis, F. The application of machine learning techniques to erosion modelling. In *Proceedings of the Third International Conference on Integrating GIS and Environmental modelling*. National Center for Geographic Information and Analysis, January 1996.

Fayyad, U., G. Piatetsky-Shapiro, and S. P. Advances in knowledge discovery and data mining. In *Data Mining to Knowledge Discovery: an Overview*, pages 1–34. American Association for Artificial Intelligence, 1996a.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases (a survey). *AI Magazine*, 3(17):37–54, 1996b.

Frank, E., L. Trigg, G. Holmes, and I. Whitten. Naive bayes for regression. *Machine Learning*, 1(41):5–26, October 2000.

Gatts, C., A. Ovalle, and C. Silva. Neural pattern recognition and multivariate data: Water typology of the Paraiba do Sul River, Brazil. *Environmental Modelling and Software*, 20(7):883–889, July 2005.

Gibbs, M., N. Morgan, H. Maier, H. M. Dandy, GC, and J. Nixon. Use of artificial neural networks for modelling chlorine residuals in water distribution systems. In *MODSIM 2003: Proceedings of the 2003 International Congress on Modelling and Simulation*, pages 789–794, 2003.

Gibert, K., R. Annicchiarico, U. Cortés, and C. Caltagirone. *Knowledge Discovery on Functional Disabilities: Clustering Based on Rules Versus Other Approaches*. IOS Press, 2005a.

Gibert, K., X. Flores, I. Rodríguez-Roda, and M. Sànchez-Marrè. Knowledge discovery in environmental data bases using GESCONDA. In *Proceedings of IEMSS 2004: International Environmental Modelling and Software Society Conference, Osnabruck, Germany*, 2004.

Gibert, K., R. Nonell, V. JM, and C. MM. Knowledge discovery with clustering: Impact of metrics and reporting phase by using klass. *Neural Network World*, pages 319–326, 2005b.

Gibert, K., M. Sànchez Marrè, and X. Flores. Cluster discovery in environmental databases using gesconda: The added value of comparisons. *AI Communications*, 4(18):319–331, 2005c.

Gibert, K. and Z. Sonicki. Clustering based on rules and medical research. *Journal on Applied Stochastic Models in Business and Industry, formerly JASMDA*, 15(4):319–324, Oct–Dec 1999.

Guariso, G. and H. Werthner. *Environmental Decision Support Systems*. Ellis Horwood-Wiley, New York, 1989.

Guo, Q., M. Kelly, and C. Graham. Support vector machines for predicting distribution of sudden oak death in california. *Ecological Modelling*, 182(1):75–90, 2005.

Hall, M. Feature selection for discrete and numeric class machine learning. Technical report, Department of Computer Science, University of Waikato, April 1999. Working Paper 99/4.

Han, J. and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001.

Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.

Holmes, G., S. Cunningham, B. Dela Rue, and A. Bollen. Predicting apple bruising using machine learning. In *Proceedings of the Model-IT Conference, Acta Horticulturae*, number 476, pages 289–296, 1998.

Keogh, E. and S. Kasetty. On the need for time series data mining benchmarks: a survey and empiricial demonstration. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 102–111. ACM Press, 2002.

Kolodner, J. *Case-Based Reasoning*. Morgan Kaufmann, 1993.

Kralisch, S., M. Fink, W.-A. Flügel, and C. Beckstein. Using neural network techniques to optimize agricultural land management for minimisation of nitrogen loading. In *MODSIM 2001: Proceedings of the 2001 International Congress on Modelling and Simulation*, pages 203–208, 2001.

Larose, D. *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley, 2004.

Lebart, L., A. Morineau, and K. Warwick. *Multivariate Descriptive Statistical Analysis*. Wiley, New York, USA, 1984.

Little, R. and D. Rubin. *Statistical Analysis with Missing Data*. Wiley, 1987.

Mas, J., H. Puig, J. Palacio, and A. Sosa-Lopez. Modelling deforestation using GIS and artificial neural networks. *Environmental Modelling and Software*, 19(5):461–471, May 2004.

McKenzie, N. and P. Ryan. Spatial prediction of soil properties using environmental correlation. *Geoderma*, (89):67–94, 1999.

Michalski, R. and R. Chilausky. Learning by being told and learning by examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. *International Journal of Policy Analysis and Information Systems*, 4(2):125–161, 1980.

Michalski, R. and R. Stepp. Learning from observation: Conceptual clustering. In Michalski, R., namesleft>1 Carbonell, J.,numnames>2 and lastname Mitchell, T., editors, *Machine Learning, an Artificial Intelligence Approach*, pages 331–363. Morgan Kaufmann, 1983.

Molina, L., L. Belanche, and A. Nebot. Feature selection algorithms: A survey and experimental evaluation. In *ICDM 2002: Proceedings of the IEEE International Conference on Data Mining*, pages 306–313, 2002.

Moore, D. and G. McCabe. *Introduction to the practice of statistics*. WH Freeman, New York, 1993. second edition.

Mora-López, L. and R. Conejo. Qualitative reasoning model for the prediction of climatic data. In *ECAI 1998: Proceedings of the 13th European Conference on Artificial Intelligence*, pages 61–75, 1998.

Nez, H., Sànchez-Marrè, and U. Cortés. Improving similarity assessment with entropy-based local weighting. In *Lecture Notes in Artificial Intelligence, (LNAI-2689): Proceedings of the 5th International Conference on Case-Based Reasoning (ICCBR*$^2003), pages377--391. Springer-Verlag, June2003$.

Núñez, H., M. Sànchez-Marrè, U. Cortés, J. Comas, M. Martinez, I. Rodríguez-Roda, and M. Poch. A

comparative study on the use of similarity measures in case-based reasoning to improve the classification of environmental system situations. *Environmental Modelling and Software*, 19(9):809–819, 2004.

of Waikato, U. Weka machine learning project, 2005. Accessed 12/08/2005.

Olsson, G. Instrumentation, control and automation in the water industry: State-of-the-art and new challenges. In *Proceedings of The 2nd IWA Conference on Instrumentation, Control and Automation (ICA 2005)*, volume 1, pages 19–31, Busan, Korea, May 29–June 2 2005.

Parr Rud, O. *Data Mining Cookbook- Modelling data for marketing, risk, and CRM*. Wiley, 2001.

Poch, M., J. Comas, I. Rodríguez-Roda, M. Sànchez-Marrè, , and U. Cortés. Designing and building real environmental decision support systems. *Environmental Modelling Software*, (19):857–873, 2004a.

Poch, M., J. Comas, I. Rodríguez-Roda, M. Sànchez-Marrè, and U. Cortés. Ten years of experience in designing and building real environmental decision support systems. what have we learnt? *Environmental Modelling and Software*, 9(19):857–873, 2004b.

Quinlan, J. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

Recknagel, F. Applications of machine learning to ecological modelling. *Ecological Modelling*, 146 (1–3):303–310, 2001.

Reutemann, P., B. Pfahringer, and E. Frank. A toolbox for learning from relational data with propositional and multi-instance learners. In *Proceedings of the Australian Joint Conference on Artificial Intelligence*, pages 1017–1023. Springer-Verlag, 2004.

Riaño, D. Learning rules within the framework of environmental sciences. In *ECAI 1998: Proceedings of the 13th European Conference on Artificial Intelligence*, pages 151–165, 1998.

Robertson, A., S. Kirshner, and P. Smyth. Hidden markov models for modelling daily rainfall occurrence over brazil. Technical Report UCI-ICS 03-27, 2003.

Rodríguez-Roda, I., J. Comas, J. Colprim, M. Poch, M. Sànchez-Marrè, U. Cortés, J. Baeza, and J. Lafuente. A hybrid supervisory system to support wastewater treatment plant operation: Implementation and validation. *Water Science and Technology*, 4–5(45):289–297, 2002.

Rodríguez-Roda, I., J. Comas, M. Poch, M. Sànchez-Marrè, and U. Cortés. Automatic knowledge acquisition from complex processes for the development of knowledge based systems. *Industrial and Engineering Chemistry Research*, 15(40):3353–3360, 2001.

Rodríguez-Roda, I., Poch, M. Sànchez-Marrè, U. Cortés, and J. Lafuente. Consider a case-based system for control of complex processes. *Chemical Engineering Progress*, 6(95):39–48, 1999.

Rubin, D. *Multiple Imputation for Nonresponse in Surveys*. Wiley, 1987.

Sanborn, S. and B. Bledsoe. Predicting streamflow regime metrics for ungauaged streams in colorado, washington, and oregon. *Journal of Hydrology*, 2005. under review.

Shearer, C. The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(4):13–22, 2000.

Siebes, A. Data mining: What it is and how it is done. In *SEBD*, page 329, 1996.

Smith, C. and J. Spate. Gully erosion in the ben chifley catchment. 2005. In Preparation.

Sànchez-Marrè, M., U. Cortés, J. Béjar, J. de Gracia, J. Lafuente, and M. Poch. Concept information in wastewater treatment plants by means of classification techniquesl: an applied study. *Applied Intelligence*, 147(7), 1997.

Sànchez Marrè, M., U. Cortés, M. Martinez, J. Comas, and I. Rodríguez-Roda. An approach for temporal case-based reasoning: Episode-based reasoning. In *Lecture Notes in Artificial Intelligence (LNAI-3620): Proceedings of the 6th International Conference on Case-Based Reasoning (ICCBR*$^2 005), pages 465--476. Springer-Verlag, August 2005.$

Sànchez Marrè, M., U. Cortés, I. Rodríguez-Roda, and M. Poch. Using meta-cases to improve accuracy in hierarchical case retrieval. *Computación y Sistemas*, 1(4):53–63, 2000.

Sànchez-Marrè, M., K. Gibert, and I. Rodríguez-Roda. *GESCONDA: A Tool for Knowledge Discovery and Data Mining in Environmental Databases*, volume 11 of *Research on Computing Science*, pages 348–364. Centro de Investigación

en Computación, Instituto Politécnico Nacional, México DF, México, 2004.

Sokal, R. and P. Sneath. *Principles of Numerical Taxonomy*. Freeman, San Francisco, 1963.

Spate, J. Data in hydrology: Existing uses and new approaches. Australian National University, 2002. Honours thesis.

Spate, J. Modelling the relationship between streamflow and electrical conductivity in Hollin Creek, southeastern Australia. In Fazel Famili, A., namesleft>1 Kok, J.,numnames>2 and lastname Peña, J., editors, *Proceedings of the 6th International Symposium on Intelligent Data Analysis*, pages 419–440, August 2005.

Spate, J. Machine learning as a tool for investigating environmental systems, In preparation. PhD Thesis.

Spate, J., B. Croke, and A. Jakeman. Data mining in hydrology. In *MODSIM 2003: Proceedings of the 2003 International Congress on Modelling and Simulation*, pages 422–427, 2003.

Spate, J. and A. Jakeman. Review of data mining techniques and their application to environmental problems. *Environmental Modelling and Software*, Under review 2006.

Su, F., C. Zhou, V. Lyne, Y. Du, and W. Shi. A data-mining approach to determine the spatio-temporal relationship between environmental factors and fish distribution. *Ecological Modelling*, 174(4):421–431, June 2004.

Swayne, D., D. Cook, and A. Buja. Xgobi: Interactive dynamic data visualization in the x window system. *Journal of Computational and Graphical Statistics*, 7(1), 1998.

Sweeney, A., N. Beebe, and R. Cooper. Analysis of environmental factors influencing the range of anopheline mosquitoes in northern Australia using a genetic algorithm and data mining methods. *Ecological Modelling*, 2004 (submitted).

Syu, M. and B. Chen. Back-propagation neural network adaptive control of a continuous wastewater treatment process. *Industrial and Engineering Chemistry Research*, 9(37), 1998.

Ter Braak, C., H. Hoijtink, W. Akkermans, and P. Verdonschot. Bayesian model-based cluster analysis of predicting macrofaunal communities. *Ecological Modelling*, 160(3):235–248, February 2003.

Vellido, A., J. Marti, I. Comas, I. Rodríguez-Roda, and F. Sabater. Exploring the ecological status of human altered streams through generative topogrpahic mapping. *Environmental Modelling and Software*, submitted 2005.

Voss, H., namesleft>1 Wachowicz, M., namesleft>1 Dzeroski, S.,numnames>2 and lastname Lanza, A., editors. *Knowledge Discovery for Environmental Management*, Knowledge-Based Services for the Public Sector Conference. KDnet, June 2004.

Wang, X., B. Chen, S. Yang, C. McGreavy, and M. Lu. Fuzzy rule generation from data for process operational decision support. *Computers and Chemical Engineering*, 21(1001), S661–S666, 1997.

Ward, J. *Hierarchical Grouping to Optimize an Objective Function*. 1963.

Weiss, G. and F. Provost. The effect of class distribution on classier learning: An empirical study. Technical report, Department of Computer Science, Rutgers University, August 2001. Technical Report ML-TR-44.

Whigham, P. Induction of a marsupial density model using genetic programming and spatial relationships. *Ecological Modelling*, 131:299–317, July 2000.

Whitten, I. and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers, 1991.

Witten, I. and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005. 2nd edition.

Wnek, J. and R. Michalski. Hypothesis-driven constructive induction in aQ17: A method and experiments. In *Proceedings of the IJCAI-91 Workshop on Evaluating and Changing Representation in Machine Learning*, pages 13–22, 1991.

Yeates, S. and K. Thomson. Applications of machine learning on two agricultural datasets. In *Proceedings of the New Zealand Conference of Postgraduate Students in Engineering and Technology*, pages 495–496, 1996.

Zhu, X. and A. Simpson. Expert system for water treatment plant operation. *Journal of Environmental Engineering*, pages 822—-829, 1996.

Zoppou, C., O. Neilsen, and L. Zhang. Regionalization of daily stream flow in Australia using wavelets and k-means. Technical report,

Australian National University, 2002. Accessed 15/10/2002.